从语言理解到多模态智能

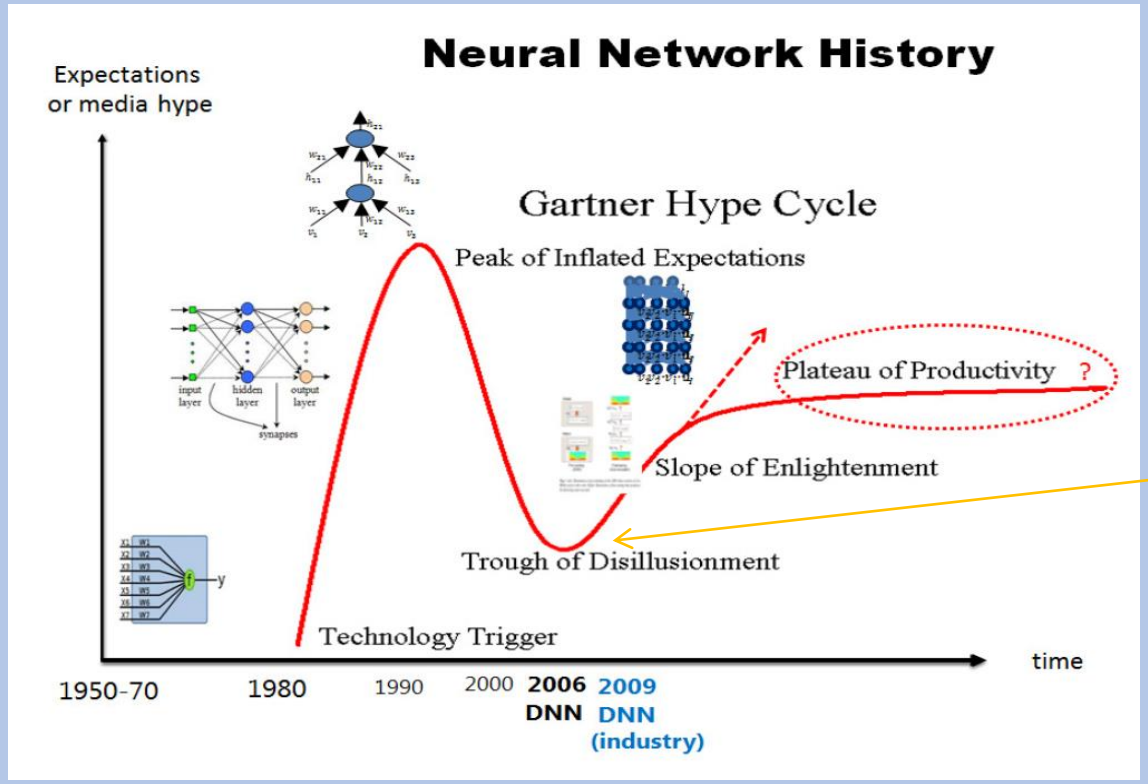
何晓冬

IEEE Fellow
京东人工智能研究院常务副院长
深度学习和语音及语言实验室主任

提纲

- 近来的一些进展
  - 深度学习
  - 语音和自然语言处理
- 语言+视觉多模态智能
  - 图像描述（Image-to-text Captioning）
  - 视觉问答（Visual Question Answering）
  - 基于文字描述合成图像（Text-to-image Synthesis）

深度学习的发展



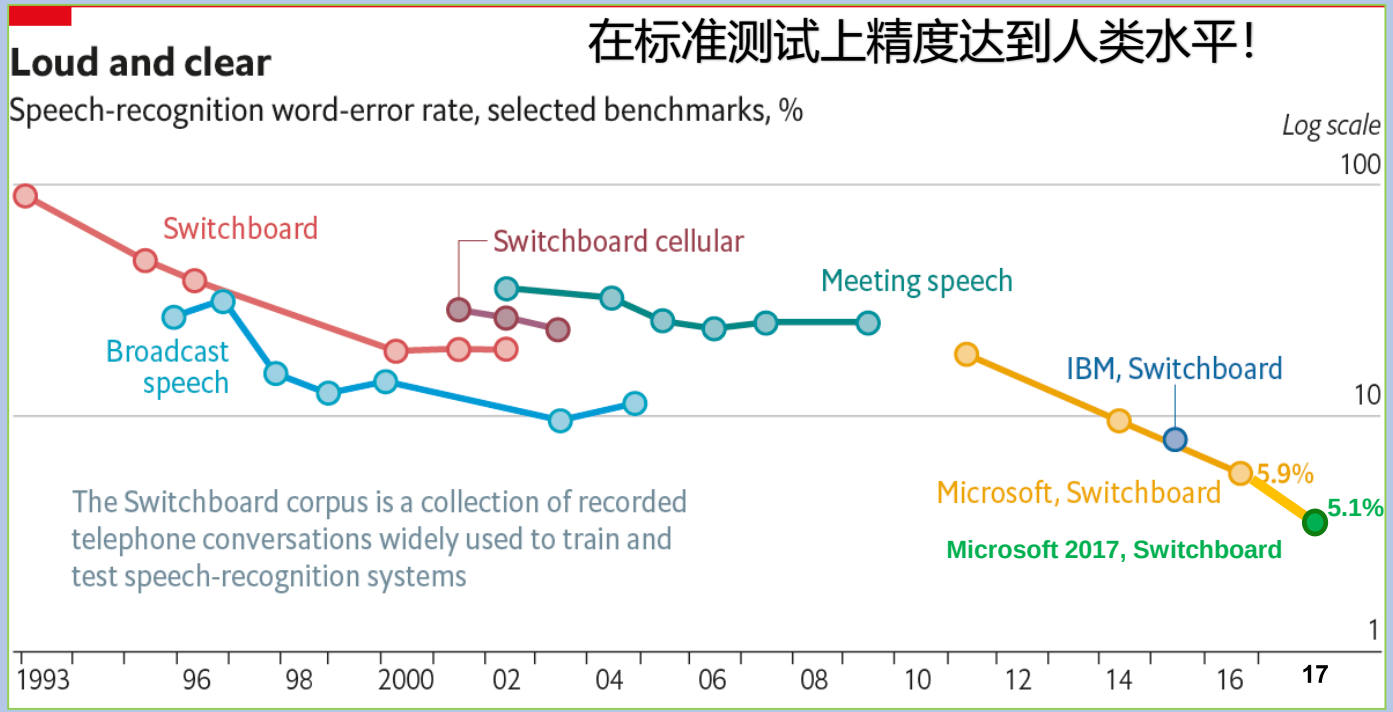
Geoff Hinton

Yoshua Bengio

Yann LeCun

2018

语音识别


在标准测试上精度达到人类水平！
Loud and clear
Speech-recognition word-error rate, selected benchmarks, %
Log scale
100
10
1
Switchboard
Switchboard cellular
Meeting speech
Broadcast speech
IBM, Switchboard
Microsoft, Switchboard
5.9%
5.1%
Microsoft 2017, Switchboard
The Switchboard corpus is a collection of recorded telephone conversations widely used to train and test speech-recognition systems
1993
96
98
2000
02
04
06
08
10
12
14
16
17

同声翻译

The universal translator on “Star Trek” comes true…

The New York Times

Scientists See Promise in Deep-Learning Programs (John Markoff November 23, 2012)

Rick Rashid in **Tianjin, China**, October, 25, 2012

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft’s top scientist, into Chinese.

2018

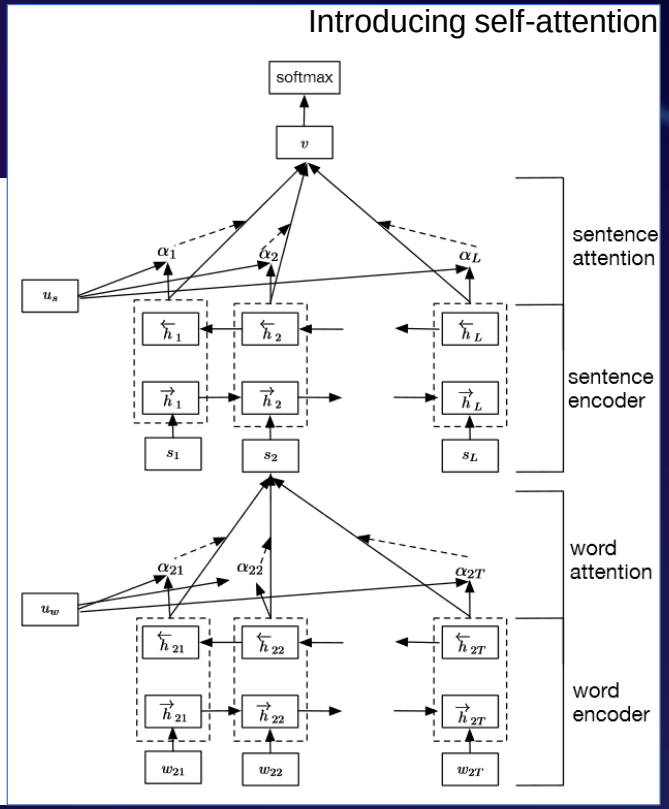
语言理解/意图分类

Hierarchical Attention Net (HAN)

我们于2016年提出层次化注意力模型 (HAN)，以更好地在词、句子、段落、等多个层面来理解语言，判断意图，并通过对神经元激活的可视化来给出一定程度的可解释性。

GT: 4 Prediction: 4

烤猪，最好吃了
带子?
我不
喜欢
带子。
这里的鸡尾酒令人惊叹，
有趣，味道好
下次我再来这个城市时，
我一定会再来一次
超推荐



【Yang, Yang, Dyer, He, Smola, Hovy, "HAN", NAACL2016】

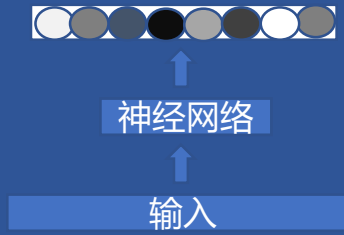
2018

语言理解/语义的表征

从自然语言中抽取出语义并将其投影到语义空间以帮助搜索、推荐、分类、问答等应用

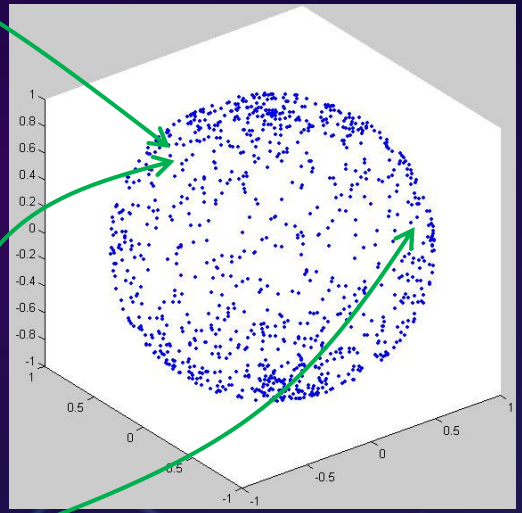
抽象的语义表征

通过深度神经网络逐步抽取语义上的不变性 (invariance)



自然语言的描述 *“小明快递了一袋苹果给外公”*

语义空间



语义相似的描述 *“外公从小明那收到了袋红富士”*

语义不同的描述 *“小明送给女友最新一代的苹果 X”*

【”DSSM”, Huang, He, Gao, Deng, Acero, Heck, CIKM2013】

2018

机器阅读理解(MRC)

机器阅读文本，回答问题

文章：印尼雅加达（CNN） - 据报道，五名欧洲人周六在印度尼西亚潜水旅行出错后获救，他们不得不在躲开科莫多巨龙的同时等待被发现。 星期四失踪后，该团体在科莫多国家公园附近的Rinca岛上的Mantaolan被发现。一同前往的潜水员向导报告共有三名英国人，一名法国人和一名瑞典人在离岛上度过了两个晚上，而这里是大型科莫多巨龙的所在地。巡逻队之后发现了他们。这个团体在周六获救。

问题: 有多少英国人获救?

答案: 三位

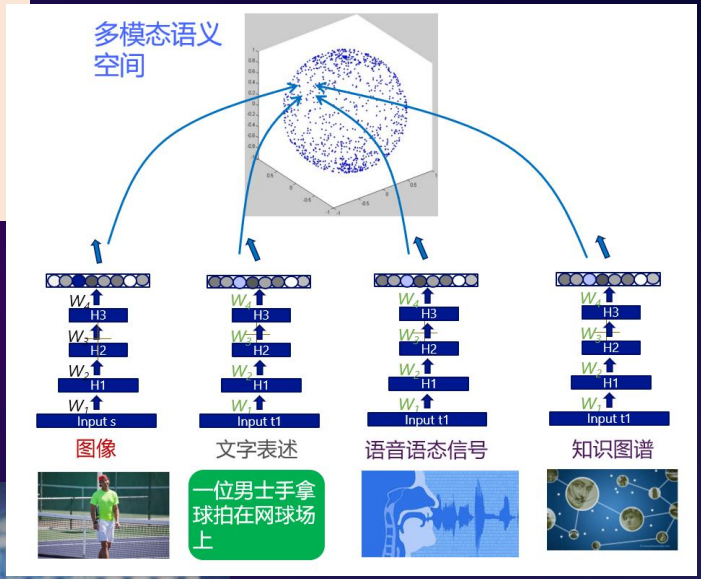
SQuAD

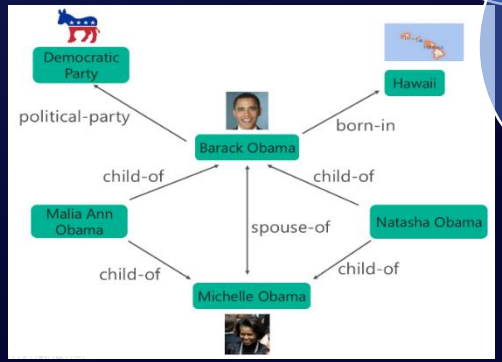
The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

多模态智能: 文本,语音,图像,知识 +

贝拉克·侯赛因·奥巴马，美国民主党籍政治家，第44任美国总统，为美国历史上第一位非裔美国人（美国黑人）总统。





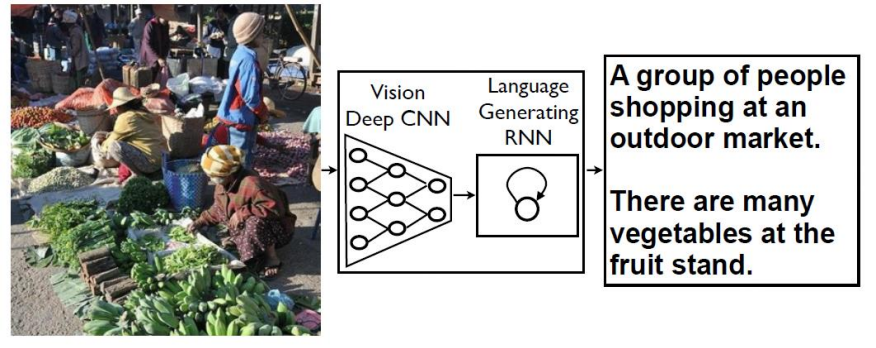
语言-视觉多模态任务

- 图像到文本描述（image-to-text / image captioning）
  - 理解图像的内容，生成自然语言来描述图像内容
- 视觉-文本问答 （visual question answering）
  - 基于对图像的理解回答相关的文本问题
- 文本到图像生成（text-to-image synthesis）
  - 基于对文字描述的理解以生成相应的图像
- 语言-视觉导航，视觉对话，跨模态信息检索 …

语言-视觉多模态：三个研究视角

- 表征预训练
  - 从原始的语言或图像信号中提取语义表征。往往是将语言和图像信号通过预训练映射到一个连续向量空间。
- 跨模态表征融合与印证
  - 在连续向量空间融合多个模态信息。往往是通过跨模态池化模型来融合语言和图像的表征（multimodal pooling），及通过注意力模型来为语言和图像的结构建立联接与印证（grounding）
- 基于多模态任务的模型优化
  - 基于特定任务设计优化目标函数及优化算法。比如包括cross-entropy，BLEU；Reinforcement learning，GAN

图像描述基础模型



End-to-End Paradigm：

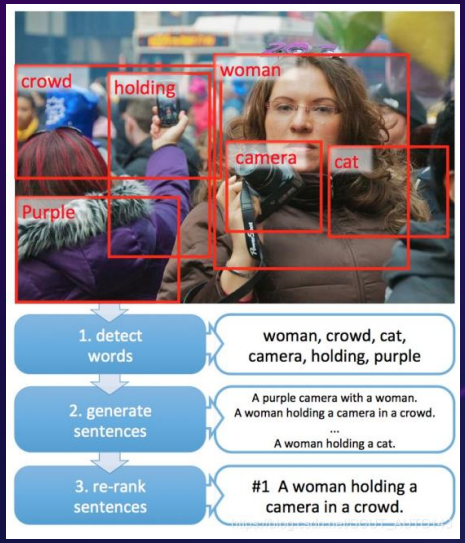
1. Encode image to vector representations（学习图像表征）
2. generate sentences（生成句子）

[Vinyals, Toshev, Bengio, Erhan， “Show and Tell: A Neural Image Caption Generator,” CVPR2015]

Cascade Paradigm：

1.detect words（检测关键物体、概念）
2.generate sentences（生成候选句子）
3.Semantic re-rank sentences（按语义表征排序）

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, et al., “From Captions to Visual Concepts and Back,” CVPR2015]

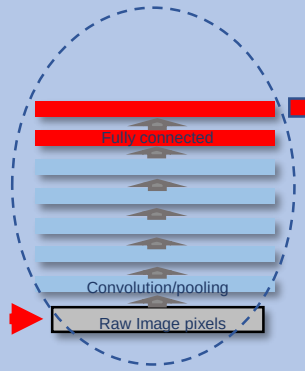


2018

建立多模态语义空间：跨模态表征学习

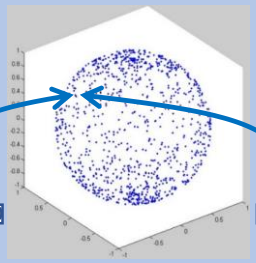
通过深度结构语义模型（DSSM）把图像和文字均表征成语义空间内的向量

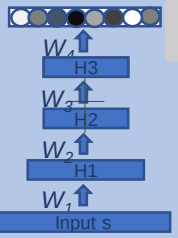
在此空间中进行语义相似度计算，生成最匹配图像内容的文字表述



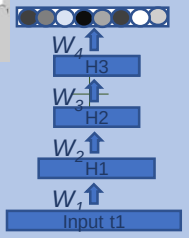
CNN

视觉-语言多模态语义空间





图像特征



文字表述: *一位男士手拿球拍在网球场上*

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, et al., “From Captions to Visual Concepts and Back,” CVPR2015]

2018

图像描述：理解场景和知识，用语言表达

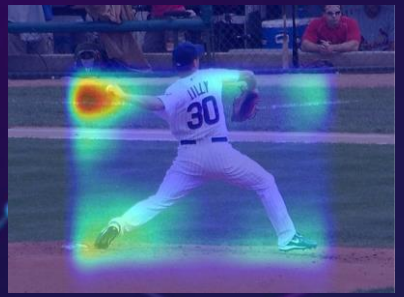
a baseball player throwing a ball

“一个棒球运动员在扔一个球。”

一个**棒球**

一个棒球**运动员**

一个棒球运动员**在扔**



一个棒球运动员在扔一个**球**

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, et al., “From Captions to Visual Concepts and Back,” CVPR2015]

2018

加入实体知识

Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.

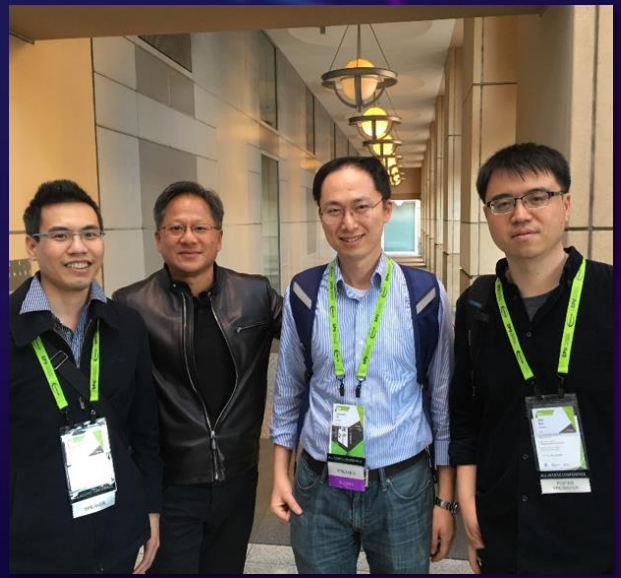
萨莎.奥巴马，玛利亚.奥巴马，米歇尔.奥巴马，彭丽媛女士等人以紫禁城为背景合影留念

2018

图片描述机器人 [http://captionbot.ai]

"A colorful bird perched on a tree branch."

一只多彩的小鸟在树枝上鸣叫。



"Jen-Hsun Huang, Xiaodong He, Jian Sun et al., that are posing for a picture."

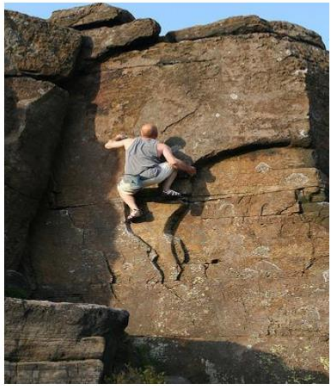
黄仁勋,何晓冬,孙剑等合影留念。

"A little boy sitting in front of a birthday cake and he seems happy." 一位小男孩坐在生日蛋糕前，看起来很高兴。

2018

表达不同的语言风格

让AI表达浪漫或者幽默的风格



CaptionBot: A man on a rocky hillside next to a stone wall.

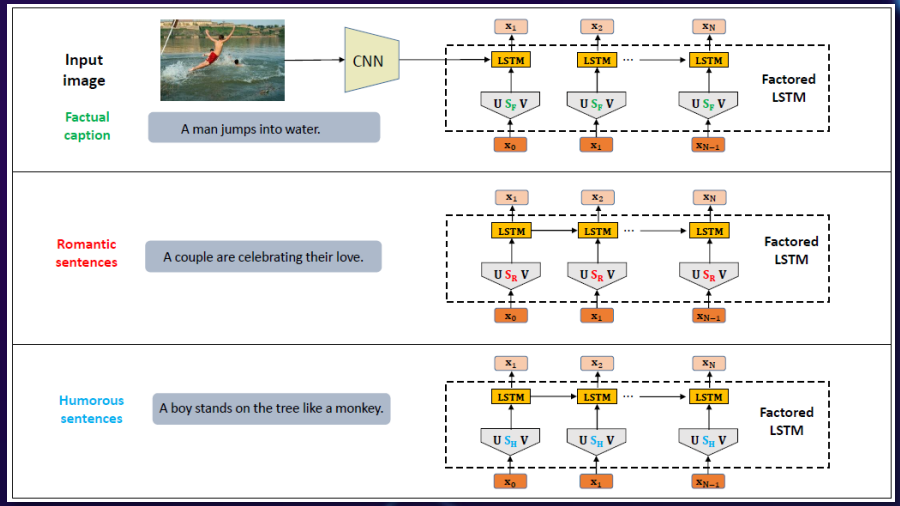
Romantic: A man uses rock climbing to overcome the obstacle in the life.

Humorous: A man is climbing the rock like a lizard.

CaptionBot: A dog runs in the grass.

Romantic: A dog runs through the grass to meet his lover.

Humorous: A dog runs through the grass in search of the missing bones.



风格神经网络模型

【Gan, Gan, He, Gao, Deng, CVPR2017】

2018

生成带情感的语言

让AI在语言表达中加入情感

分类: 户外，女士
语义: 一位穿着蓝色T恤的女士
情感+: 美丽得像一位天使!

分类: 女士，小狗
语义: 一位女士和一只狗在相机前摆姿势
情感+: 啊真可爱，我是说这只小狗耶 ☺

综合图像和语言推理，回答问题

那两把蓝色椅子之间是什么？

一把伞

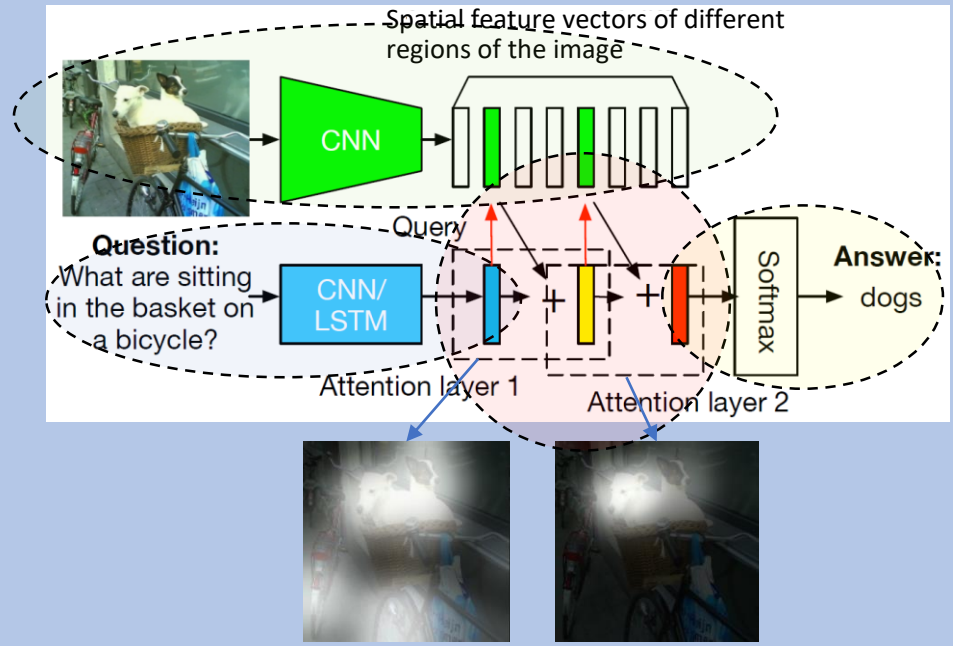
【Yang, He, Gao, Deng, Smola, CVPR2016】

2018

视觉-语言多模态推理问答

SANs perform multi-step reasoning

1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor
5. End-to-end learning using SGD



[Stacked Attention Networks, Yang, He, Gao, Deng, Smola, CVPR 2016]

提取视觉表征

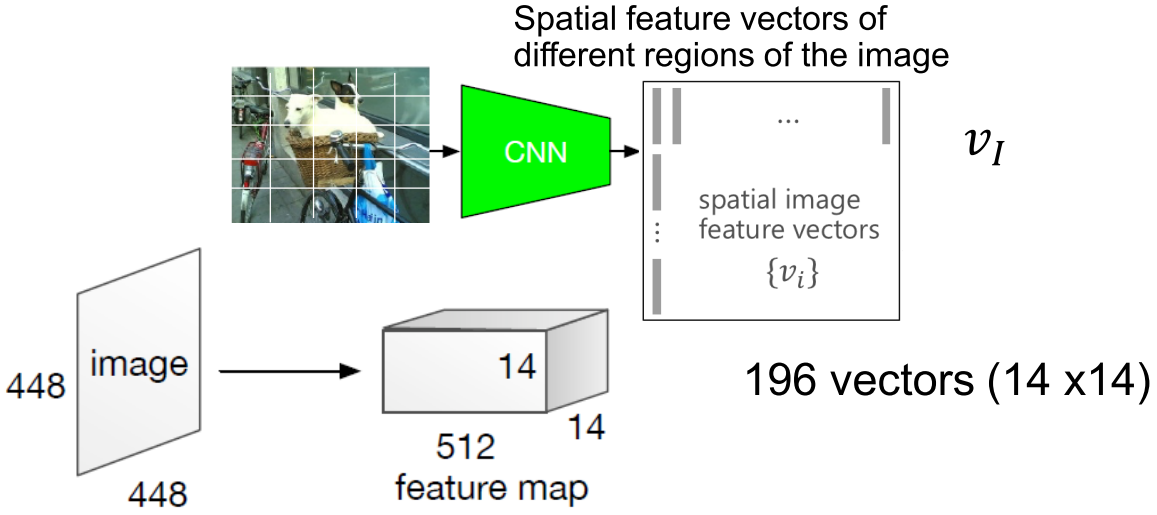
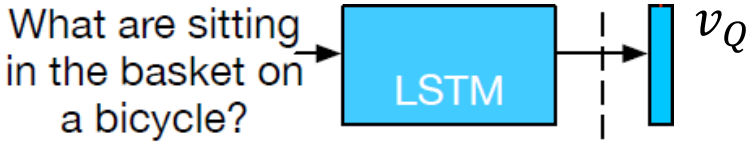


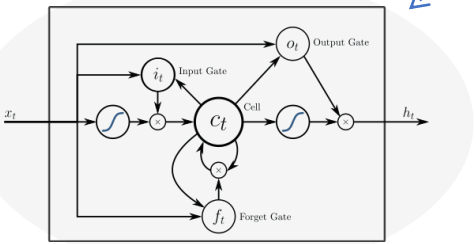
Figure 2: CNN based image model

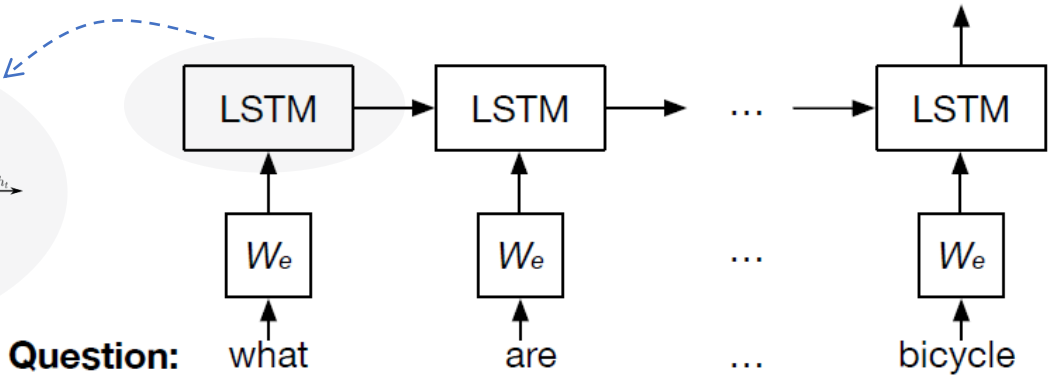
$$f_I = \text{CNN}_{vgg}(I). \quad v_I = \tanh(W_I f_I + b_I)$$

提取语言表征

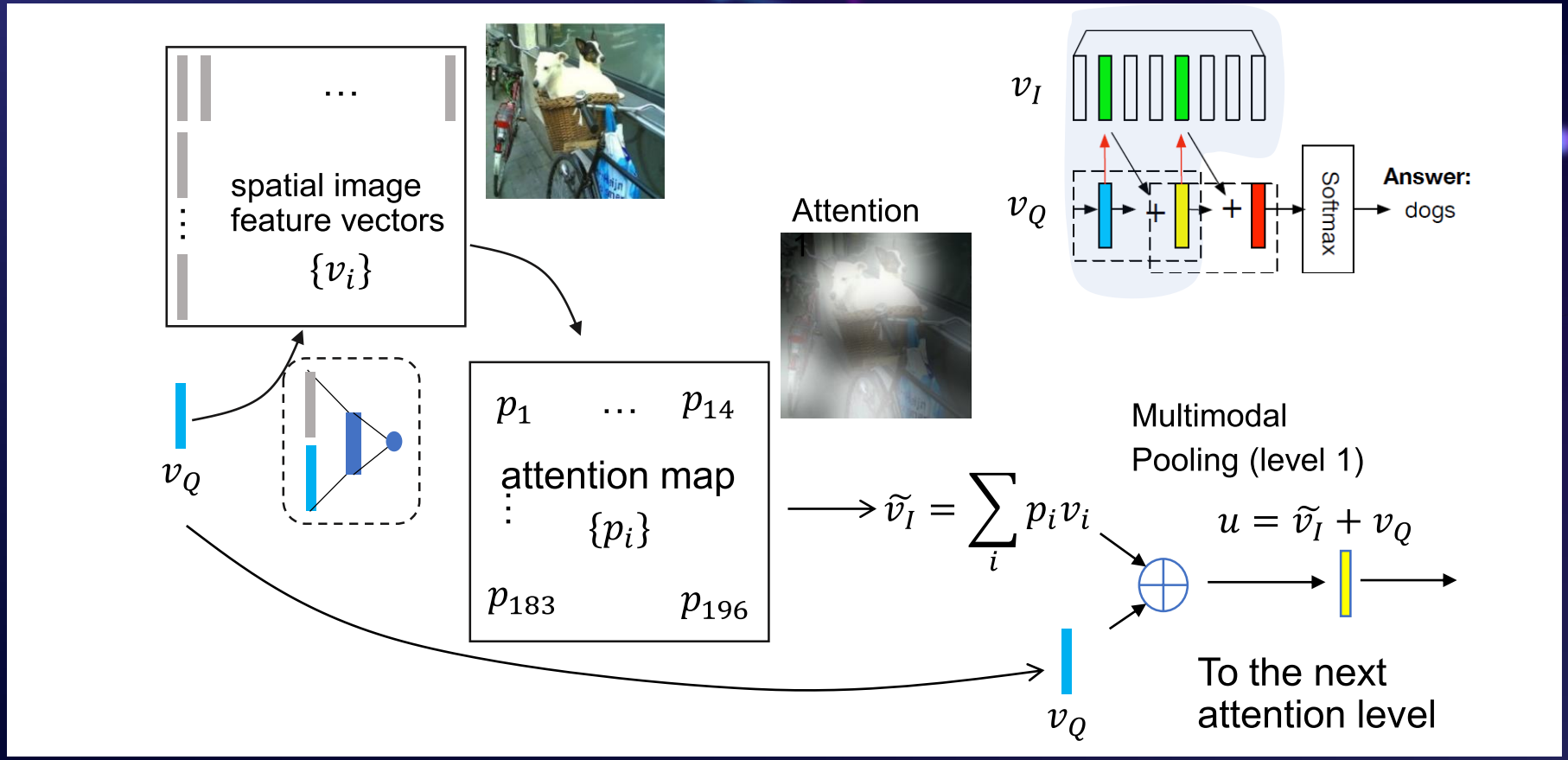


A LSTM cell

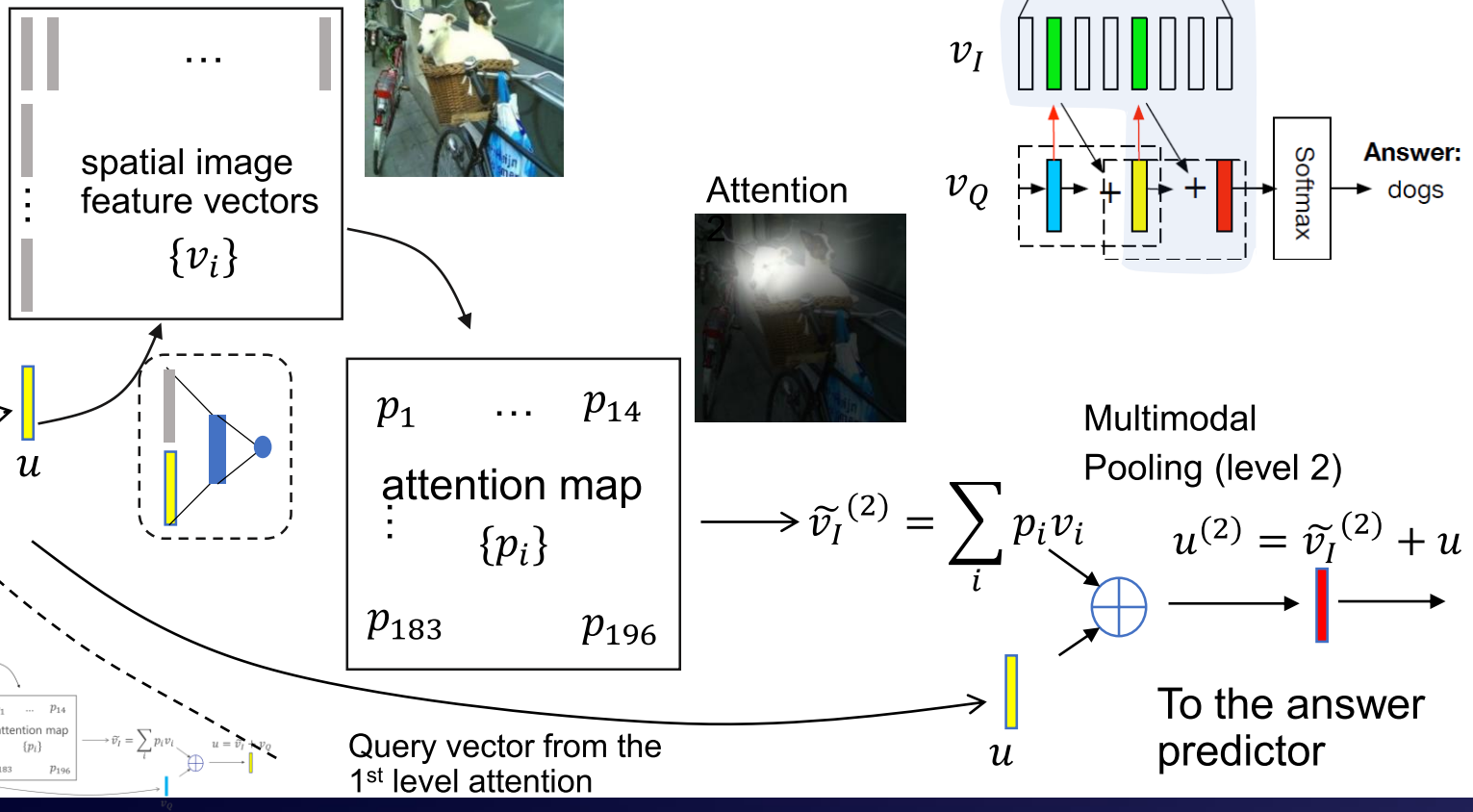




跨模态表征融合与印证(pooling & grounding)



跨模态表征融合与印证


spatial image feature vectors
$\{v_i\}$
attention map
$\{p_i\}$
p_1 … p_{14}
p_{183} p_{196}
u
Attention
v_I
v_Q
Softmax
Answer:
dogs
Multimodal Pooling (level 2)
$\tilde{v}_I^{(2)} = \sum_i p_i v_i$
$u^{(2)} = \tilde{v}_I^{(2)} + u$
To the answer predictor
Query vector from the 1st level attention

Bottom-Up and Top-Down Attention

注意力模型的一个新视角

In human visual system, there are two kinds of attentions:

Top-down attention:

proactively initiated by the current task (e.g., look for something)

Bottom-up attention:

spontaneously emerge from visual salient stimuli

2018

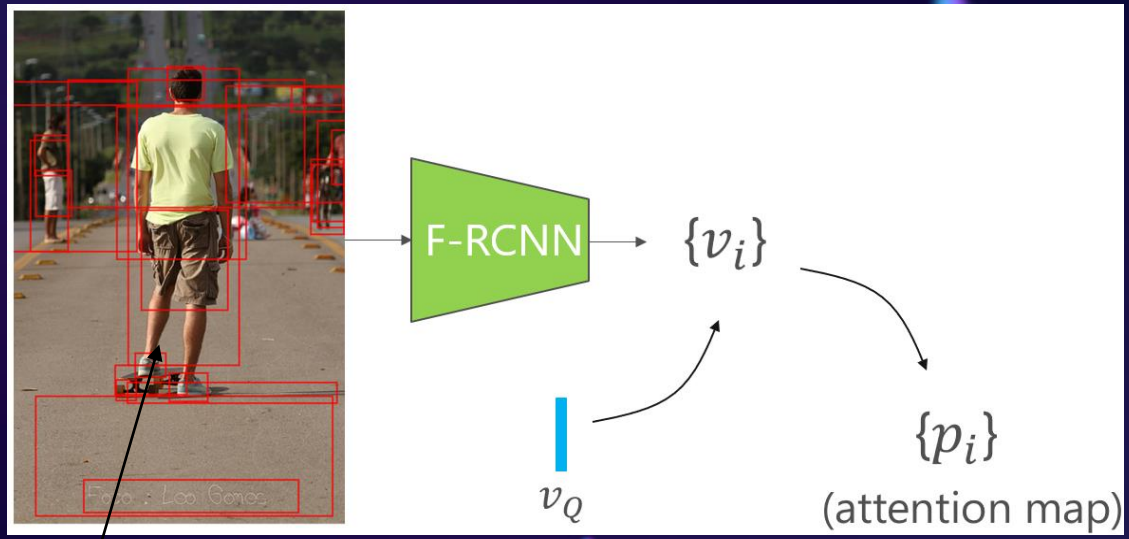
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]*, **Xiaodong He**[2], **Chris Buehler**[2], **Damien Teney**[3]
Mark Johnson[4], **Stephen Gould**[1], **Lei Zhang**[2]
[1]Australian National University [2]Microsoft Research
[3]University of Adelaide [4]Macquarie University

Bottom-Up attention mechanism (new)

Bottom-Up attention:

- Use F-RCNN to detect key objects
- Compute spatial feature vector for each object
- Keep complete visual information for each object



Attend on actual objects, rather than on uniform grid regions like conventional top-down attention

Combine Bottom-Up & Top-Down Attention

Adopt similar terminology to humans' attention system:

- attention mechanisms driven by non visual or task-specific context as 'top-down'
- purely visual feed-forward attention mechanisms as 'bottom-up'.

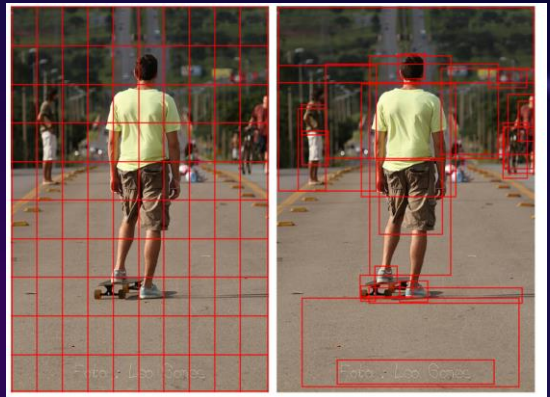
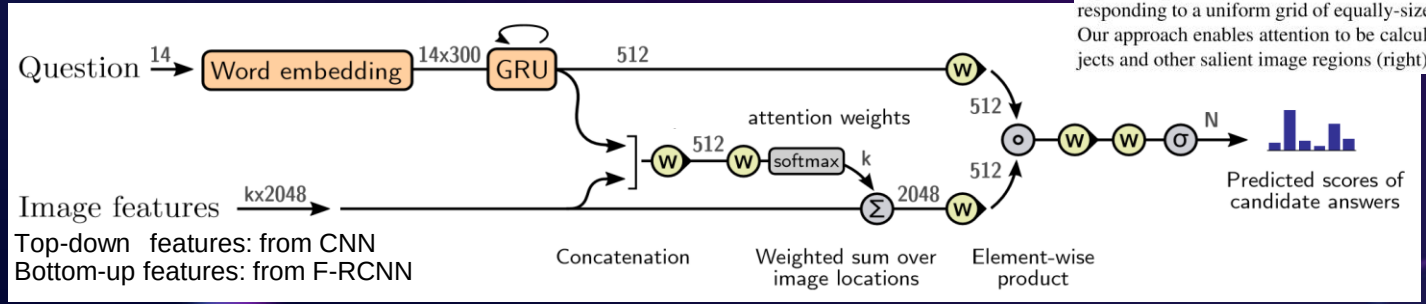


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Overall Attention Net for VQA:



Top-down features: from CNN
Bottom-up features: from F-RCNN

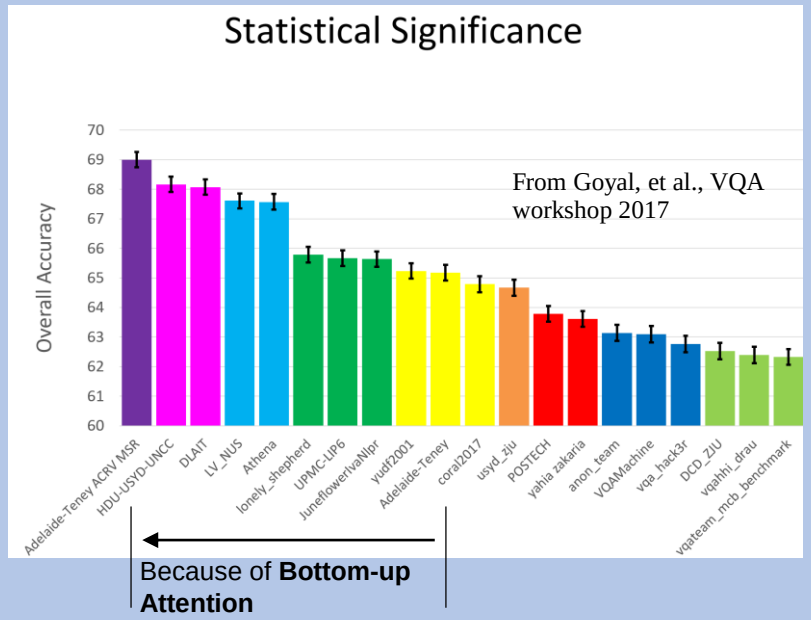
2018

Attention Example

Question: What room are they in? Answer: kitchen

Figure 6. VQA example illustrating attention output. Given the question 'What room are they in?', the model focuses on the stove-top, generating the answer 'kitchen'.

VQA Challenge @ CVPR2017

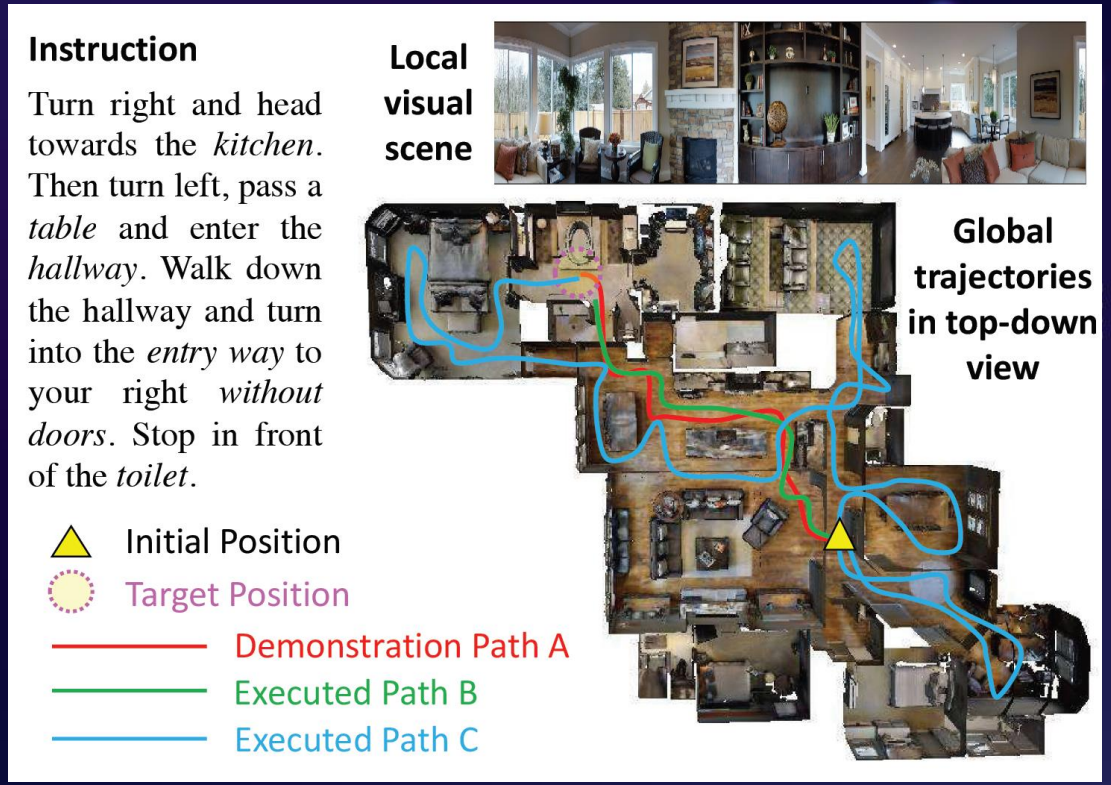


[1] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR18
[2] Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge, CVPR18

之后, 几乎所有的VQA队伍都使用了“Bottom-Up and Top-Down (BUTD)”注意力模型或其变种。

视觉–语言多模态导航

结合语言理解和对环境的视觉信息建模，智能代理能按指令从一个地方走到另一个地方



[Anderson et al., CVPR2018; Wang et al., CVPR 2019]

理解语言，用绘画来表达

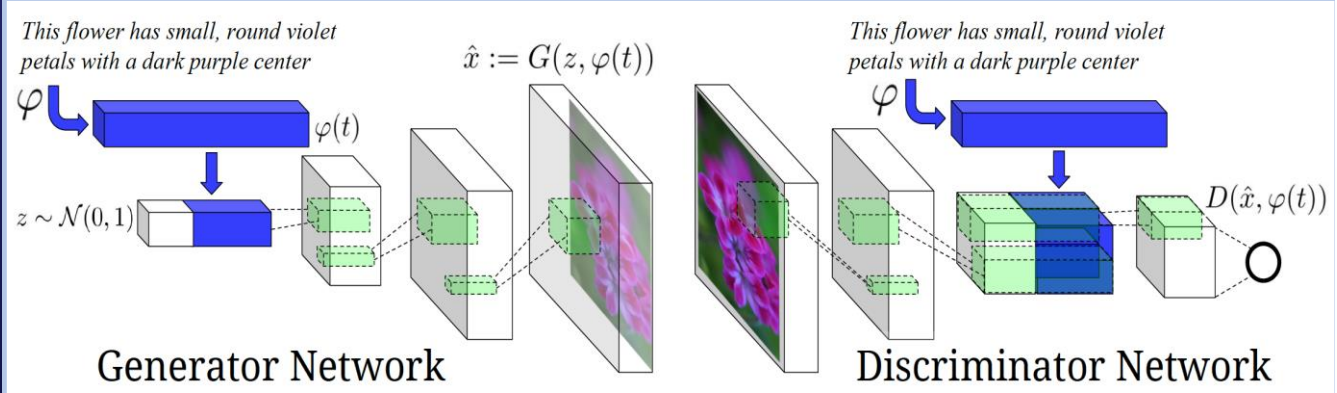


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

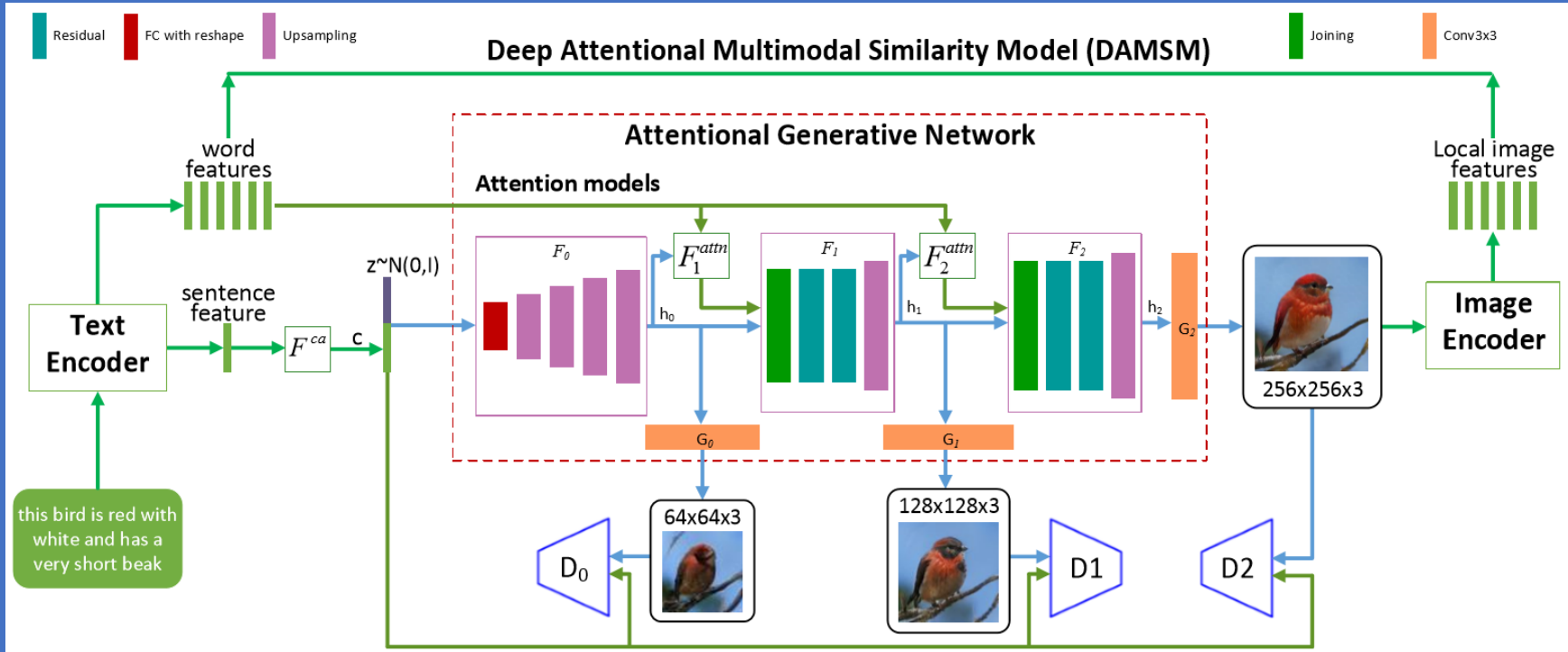
this small bird has a pink breast and crown, and black primaries and secondaries.

Objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

[Reed et al., Generative adversarial text-to-image synthesis, ICML, 2016]

AttnGAN: 智能绘画机器人



The final objective function: $L = L_{GAN} + \lambda L_{DAMSM}$

【Xu, Zhang, Huang, Zhang, Gan, Huang, He, “AttnGAN,” CVPR2018】

理解语言，用绘画来表达

一只红羽毛白肚子的短咀小鸟

2018

【Xu, Zhang, Huang, Zhang, Gan, Huang, He, “AttnGAN,” CVPR2018】

this bird has wings that are blue and has a red belly

this bird has a green crown black primaries and a white belly

this bird has a yellow crown and a black eyering that is round

a small red and white bird with a small curved beak

飞速发展的技术领域

[Reed et al., Generative adversarial text-to-image synthesis, ICML, **2016**]

[Xu et al., AttnGAN: Fine-grained text to image generation with Attentional GANs, CVPR **2018**]

[Brock et al., BigGAN: A New State of the Art in Image Synthesis, ICLR **2019**]

2018

AI+Art

From Knowledge Map to Mind Map: Artificial Imagination

Ruixue Liu*§, Baoyang Chen†§, Xiaoyu Guo*, Yan Dai*, Meng Chen*, Zhijie Qiu†, Xiaodong He‡
*JD AI Platform & Research, Beijing, China
†Faculty of School of Experimental Arts, Central Academy of Fine Arts, Beijing, China
‡JD AI Research, Beijing, China
{liuruixue,guoxiaoyu5,daiyan5,chenmeng20,xiaodong.he}@jd.com, {chenbaoyang, qiuzhijie}@cafa.edu.cn

Imagination is one of the most important factors which makes an artistic painting unique and impressive. we propose a novel approach to inject rich imagination into a special painting art Mind Map creation, and finally apply Dadaism and impossibility of improvisation principles into painting process…

Figure 1: An example of Mind Map

2018

【Liu, et al., 2019】

多模态智能的下一步

- 融合多个子领域（语言、视觉、语音、知识工程…）
- 跨模态预训练、概念/实体grounding、神经-符号处理机
- 通过复杂跨模态问题驱动基础研究
- 打造更成熟和多元化的实际多模态智能应用

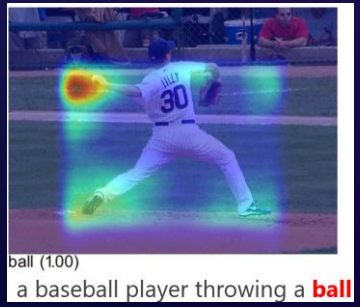
2018

跨越语言和视觉的理解和表达

Universal Chatbot, Digital Assistant, Mixed Reality, …

Multimodal Intelligence: perception, reasoning, and expression across language & vision

Image-to-language



Visual QA/Dialog

Q: what are sitting in the basket on a bicycle?
A: dogs.

Language-to-image

This bird is red with white and has a very short beak

Deep Representation Learning / Deep Attention Mechanisms

2018

THANKS

2018

xiaodong.he@jd.com

智汇京东 · 开放共赢